OCCUPATION MEASURES IN AVERAGE COST MARKOV DECISION PROCESSES(Optimization Theory and its Applications in Mathematical Systems)

AUTHOR(S):

SOH, KINGESTU; HOSAKA, MASANORI

数理解析研究所講究録
第 899 巻 1995 年 25-33

OCCUPATION MEASURES IN AVERAGE COST MARKOV DECISION PROCESSES

千葉大学 理学研究科 宋 金杰(KINGESTU SOH)
千葉大学 理学研究科 保坂正徳(MASANORI HOSAKA)

ABSTRACT. We consider the average cost Markov decision processes (MDP's) with general state and action spaces. Extending the idea in Borkar's excellent paper [3,4]. We define an extended occupation measure associated with the class of policies for MDP's and an annexed index (called a power), by which the validity for optimization is measured. Also, by construction of an extended occupation measure, the policy with robustness for the cost function is given. The proofs are done without continuity and compactness and universally and/or analytically measurable policies are unnecessary to describe the results, which are new in this paper.

1. INTRODUCTION AND NOTATION

Studies on Markov decision processes(MDP's) are done mainly in the case of the known cost function (see [9]). But in many applicable areas, it often occurs that the cost function is unknown or partially unknown. In such a case, the policy with robustness for the cost function will be useful. In this paper, we consider average cost MDP's with general state and action spaces and try to construct the policy robust for the cost function.

For the sake of this purpose, extending the idea of occupation measures in Borker [3,4], we introduce an extended one associated with the class of policies for MDP's and an annexed index (called a power), by which the validity for optimization is measured. Also, constructing the occupation measure by the method of obtaining a measure from a pre-measure (see [7,8]), the policy with robustness for the cost function is given. The discussion in this paper is done under some minorization conditions which is often used in the study of ergodicity of Markov chains ([10]).

The case of general state and action spaces is usually discussed under analytic or universal measurability (for example, see [1,6]). But, in this paper, the proofs are done only under Borel measurability. Also, the hypotheses of continuity and compactness are excluded from our discussion. These fact are new as far as we are aware.

A Borel set is a Borel subset of some complete separable metric space. For any Borel set X and Y, let $\mathcal{B}_X$ be the set of all Borel subsets of X, $\mathcal{P}(X)$ and $B_+(X)$ the sets of all probability measures and non-negative real valued and bounded Borel measurable functions on X respectively, and $T(X|Y)$ the set of all stochastic kernels on $\mathcal{B}_X \times Y$, i.e., $q \in T(X|Y)$ means that for each $y \in Y$, $q(\cdot|y)$ ia a probability measure on $\mathcal{B}_X$ and for each $D \in \mathcal{B}_X$, $q(D|\cdot)$ is a Borel measurable function on Y.

Let (S, A, c, Q) be MDP's, where S and A are Borel sets, $c \in B_+(S \times A)$ and $Q \in T(S|S \times A)$. The state of the process is denoted as a point in S. If, in state $x \in S$, we take action $a \in A$, the process incurs a cost $c(x, a)$ and moves to state x' on the next transition with the probability $Q(\cdot|x, a)$. We treat the case of the cost function being unknown, so that c is thought of as a variable. For each $t \geq 0$, let denote by X_t and Δ_t the state and action at t-th time respectively.

The sample space is the product space $\Omega = (S \times A)^\infty$, where X_t and Δ_t are projections from Ω on the t-th factors S and A.

A policy $\pi = (\pi_0, \pi_1, \cdots)$ denotes a rule of taking actions that depend on both the current state and the past history of the process and which can be randomized, so that $\pi_t \in T(A|(S \times A)^t \times S)$ for $t \geq 0$.

For any $\Phi \in T(A|S)$, if we choose the action randomly according to $\Phi(\cdot|x)$ in a current state $x \in S$, regardless of the past history, such a policy is called randomized stationary and denoted by $\Phi^{(\infty)}$.

Let denotes by $B(S \to A)$ the set of all Borel measurable functions $u : S \to A$. A randomized stationary policy $\Phi^{(\infty)}$ is called stationary if there exists an $f \in B(S \to A)$ such that $\Phi(\{f(x)\}|x) = 1$ for all $x \in S$. Such a policy will be written by $f^{(\infty)}$.

For each policy $\pi \in \Pi$ and initial state distribution $\nu \in \mathcal{P}(S)$, we can define the probability measure P_ν^π on the sample space Ω in an obvious way.

We shall consider the following average cost criterion: For any $\pi \in \Pi$, $\nu \in \mathcal{P}(S)$ and $c \in B_+(S \times A)$, let

$$\Psi(\nu, \pi, c) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\nu^\pi [c(X_t, \Delta_t)], \tag{1.1}$$

where E_ν^π is the expectation operator w.r.t. P_ν^π. Let

$$\Psi(\nu, c) := \inf_{\pi \in \Pi} \Psi(\nu, \pi, c).$$

For any subspace $B \subset B_+(S \times A)$, $\nu \in \mathcal{P}(S)$ and $\varepsilon \geq 0$, we say that $\pi^* \in \Pi$ is (ν, ε)-optimal for B, if

$$\Psi(\nu, \pi^*, c) \leq \Psi(\nu, c) + \varepsilon \quad \text{for all} \quad c \in B.$$

Also $\pi^* \in \Pi$ is ε-optimal for B, if

$$\Psi(x, \pi^*, c) \leq \Psi(x, c) + \varepsilon \quad \text{for all} \quad c \in B \text{ and } x \in S,$$

where the degenerate initial distribution concentrated at the point x is denoted by x. As a subclass of $B_+(S \times A)$ in the above, the following will be of interest :

$$B_+^M := \{c \in B_+(S \times A) \mid c(x, a) \leq M \quad \text{for all} \quad (x, a) \in S \times A\} \quad \text{for} \quad M > 0.$$

We will need the following well-known results which is used in the sequel.

Lemma 1.1 ([2]). *For any $\Phi \in T(A|S)$ and $u \in B_+(S \times A)$, there exists an $f \in B(S \to A)$ such that*

$$u(x, f(x)) \leq \int u(x, a) \Phi(da|x) \qquad \text{for all } x \in S.$$

Lemma 1.2 (Tauberian theorem, cf. [11]). *Let $\{a_t\}$ be a bounded sequence of real numbers. Then:*

$$(i) \qquad \liminf_{\beta \uparrow 1} (1 - \beta) \sum_{t=0}^{\infty} \beta^t a_t \geq \liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} a_t,$$

$$(ii) \qquad \limsup_{\beta \uparrow 1} (1 - \beta) \sum_{t=0}^{\infty} \beta^t a_t \leq \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} a_t \quad \text{and}$$

$$(iii) \qquad \text{if } \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} a_t \text{ exists}, \quad \lim_{\beta \uparrow 1} (1 - \beta) \sum_{t=0}^{\infty} \beta^t a_t = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} a_t.$$

In Section 2, an occupation measure associated with the class of policies is given and its validity for optimization is discussed under a minorization condition. In Section 3, we are concerned with the construction of an occupation measure. In Section 4, optimization is discussed in the treatment of an occupation measure.

2. Occupation measure : Definition

In this section, we define the occupation measure of the average cost case by considering it from the discounted one, and give its validity for optimization under a minorization condition.

For any $\beta(0 < \beta < 1)$, $\nu \in \mathcal{P}(S)$, $\pi \in \Pi$ and $g \in B_+(S \times A)$, let

$$L^{\pi}_{\nu,\beta}(g) := \sum_{t=0}^{\infty} \beta^t E^{\pi}_{\nu}[g(X_t, \Delta_t)] \quad \text{and} \tag{2.1}$$

$$L^{\pi}_{\nu}(g) := \liminf_{\beta \to 1} (1-\beta) L^{\pi}_{\nu,\beta}(g). \tag{2.2}$$

For any $D \in \mathcal{B}_{S \times A}$, when $g = I_D$ in (2.2), we write it simply by $L^{\pi}_{\nu}(D)$, where I_D is the indicator i.e., $I_D(x) = 1$ if $x \in D$ and $I_D(x) = 0$ if $x \notin D$.

Let $\Pi' \subset \Pi$, $\nu \in \mathcal{P}(S)$ and $\mu \in \mathcal{P}(S \times A)$. Then, if there exists a constant $K > 0$ such that

$$K \int g\mu(d(x,a)) \leq \inf_{\pi \in \Pi'} L^{\pi}_{\nu}(g) \quad \text{for all } g \in B_+(S \times A), \tag{2.3}$$

μ is called an occupation measure associated with Π' and ν. When $\Pi' = \Pi$, μ is simply called an occupation measure associated with ν. Let

$$K^*(\Pi', \nu) := \max\{K \mid K \text{ satisfies (2.3)}\}.$$

In order to prove one of the main results we need the following minorization condition which is often used in the study of ergodicity of Markov chains (for example, see [10]).

Condition A. *There exists a measure $\gamma(\cdot)$ on S such that $\gamma(S) > 0$ and $Q(D|x,a) \geq \gamma(D)$ for all $D \in \mathcal{B}_S$ and $(x,a) \in S \times A$.*

Under Condition A, the following holds (for example see [10]):

$$\Psi(\nu, \Phi^{(\infty)}, c) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E^{\Phi^{(\infty)}}_{\nu}[c(X_t, \Delta_t)] \quad \text{for all } \nu \in P(S), \Phi \in T(A|S). \tag{2.4}$$

For the rest of this section we assume that Condition A holds. Let

$$\overline{Q}(\cdot|x,a) := Q(\cdot|x,a) - \gamma(\cdot) \quad \text{for any } (x,a) \in S \times A.$$

Then, if we put $\alpha := \overline{Q}(S|x,a)$, clearly $0 \leq \alpha < 1$.

Theorem 2.1. *Suppose that Condition A holds. Then, for the occupation measure μ associated with $\Pi' \subset \Pi$, $\nu \in \mathcal{P}(S)$, there exists a randomized stationary policy $\Phi^{(\infty)}$ satisfying*

$$\Psi(\nu, \Phi^{(\infty)}, c) \leq \inf_{\pi \in \Pi'} \Psi(\nu, \pi, c) + \frac{\|c\|(1 - K^*(\Pi', \nu))}{1 - \alpha} \tag{2.5}$$

for all $c \in B_+(S \times A)$, where $\|\cdot\|$ is the supremum norm.

We observe from (2.5) that the occupation measure μ becomes as more powerful for optimization as $K^*(\Pi', \nu)$ nears to 1. So, we will call it a power of μ.

We provide a proof of Theorem 2.1 in a series of Lemmas, some of independent interests. For any $\Phi \in T(A|S)$, the t-step transition w. r. t. $\overline{Q}$ are defined by

$$\overline{Q}^{(1)}(\cdot|x,\Phi) = \int \overline{Q}(\cdot|x,a)\Phi(da|x) \quad \text{and} \tag{2.6}$$

$$\overline{Q}^{(t+1)}(\cdot|x,\Phi) = \int \overline{Q}^{(t)}(\cdot|x_1,\Phi)\overline{Q}^{(1)}(dx_1|x,\Phi) \quad (t \geq 0) \tag{2.7}$$

$$\text{where } \overline{Q}^{(0)}(D|x,\Phi) = I_D(x) \quad \text{for all } D \in \mathcal{B}_S.$$

We have the following lemma:

Lemma 2.1. *For any $\Phi \in T(A|S), x \in S$ and $t \geq 0$,*

(*i*) $$\overline{Q}^{(t)}(S|x,\Phi) = \alpha^t \quad \textit{and}$$

(*ii*) $$E_x^{\Phi(\infty)}[c(X_t,\Delta_t)] = \int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi) + B_t \quad , \textit{where}$$

$$B_t := \sum_{k=0}^{t-1} \int\int c(z,\Phi)\overline{Q}^{(k)}(dz|x,\Phi)\gamma(dx), \quad c(z,\Phi) = \int c(z,a)\Phi(da|z).$$

Proof. The proof proceeds by induction. For $t = 0, 1$, (i) is obviously true. Suppose that (i) holds for $t \geq 0$. By (2.7),

$$\overline{Q}^{(t+1)}(S|x,\Phi) = \alpha^t \int \overline{Q}^{(1)}(dx_1|x,\Phi) = \alpha^{t+1},$$

which shows that (i) holds for $t+1$. Let $\mathcal{B}_t := \mathcal{B}(X_0,\Delta_0,\cdots,X_{t-1},\Delta_{t-1},X_t)$ and $\mathcal{B}'_t := \mathcal{B}(X_0,\Delta_0,\cdots,X_t,\Delta_t)$, where $\mathcal{B}(Z)$ is the sub-σ-field generated by the random element Z. For simplicity, put $E := E_x^{\Phi(\infty)}$. Then, clearly it holds that

$$\begin{aligned} E[c(X_1,\Delta_1)] &= E[\int c(z,\Phi)\overline{Q}(dz|X_0,\Delta_0)] + \int c(z,\Phi)\gamma(dz) \\ &= \int c(z,\Phi)\overline{Q}^{(1)}(dz|x,\Phi) + B_1, \end{aligned}$$

which implies that (ii) holds for $t = 1$. Suppose that (ii) holds for t.

$$\begin{aligned} E[c(X_{t+1},\Delta_{t+1})] &= E[E[c(X_{t+1},\Delta_{t+1})|\mathcal{B}_{t+1}]] = E[E[c(X_{t+1},\Phi)|\mathcal{B}'_t]] \\ &= E[\int c(z,\Phi)\overline{Q}(dz|X_t,\Delta_t)] + \int c(z,\Phi)\gamma(dz) \\ &= E[\overline{c}(X_t,\Delta_t)] + \int c(z,\Phi)\gamma(dz), \end{aligned}$$

$$\text{where} \quad \overline{c}(x,a) = \int c(z,\Phi)\overline{Q}(dz|x,a).$$

Here, using the inductive assumption and (2.7), we get

$$\begin{aligned} &E[c(X_{t+1},\Delta_{t+1})] \\ &= \int \overline{c}(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi) + \sum_{k=0}^{t-1}\int \overline{c}(z,\Phi)\overline{Q}^{(k)}(dz|y,\Phi)\gamma(dy) + \int c(z,\Phi)\gamma(dz) \\ &= \int c(z,\Phi)\overline{Q}^{(t+1)}(dz|x,\Phi) + \sum_{k=1}^{t}\int c(z,\Phi)\overline{Q}^{(k)}(dz|y,\Phi)\gamma(dy) + \int c(z,\Phi)\gamma(dz) \\ &= \int c(z,\Phi)\overline{Q}^{(t+1)}(dz|x,\Phi) + B_{t+1} \end{aligned}$$

This shows that (ii) holds for $t+1$. □

Let the discounted total cost $h_\beta(x,\Phi^{(\infty)})$ with $\beta (0 \leq \beta < 1)$ be

$$h_\beta(x,\Phi^{(\infty)}) := \sum_{t=0}^{\infty} \beta^t E_x^{\Phi(\infty)}[c(X_t,\Delta_t)], \quad \text{for each } x \in S, \Phi \in T(A|S).$$

By Lemma 2.1(ii), h_β can be rewritten as follows:

(2.8) $$h_\beta(x,\Phi^{(\infty)}) = \sum_{t=0}^{\infty} \beta^t \int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi) + \sum_{t=0}^{\infty} \beta^t B_t.$$

In order to argue the average cost case from the discounted cost one, we define the following functions with suppression of Φ:

$$h_\beta(y,x) := h_\beta(y,\Phi^{(\infty)}) - h_\beta(x,\Phi^{(\infty)}), \tag{2.9}$$

$$h(x) := \sum_{t=0}^{\infty} \int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi), \tag{2.10}$$

$$h(y,x) := h(y) - h(x). \tag{2.11}$$

By Lemma 2.1(i), $\sup_{x\in S} |\int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi)| \le ||c||\alpha^t, (t \ge 0)$, so that $h(x)$ is well-defined, where $||\cdot||$ is the supremum norm.

Lemma 2.2. *For any $\Phi \in T(A|S)$,*

$$\lim_{\beta\uparrow 1} ||h_\beta(\cdot,\cdot) - h(\cdot,\cdot)|| = 0 \quad \textit{and} \tag{i}$$

$$\lim_{\beta\uparrow 1} ||(1-\beta)h_\beta(\cdot,\Phi^{(\infty)}) - \Psi(\cdot,\Phi^{(\infty)},c)|| = 0. \tag{ii}$$

Proof. For (i), by (2.8) we get

$$h_\beta(y,x) = \sum_{t=0}^{\infty} \beta^t \int c(z,\Phi)\overline{Q}^{(t)}(dz|y,\Phi) - \sum_{t=0}^{\infty} \beta^t \int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi),$$

so that, from Lemma 2.1(i),

$$\begin{aligned} &|h_\beta(y,x) - h(y,x)| \\ &\le \sum_{t=0}^{T}(1-\beta^t)[\int c(z,\Phi)\overline{Q}^{(t)}(dz|y,\Phi) + \int c(z,\Phi)\overline{Q}^{(t)}(dz|x,\Phi)] + \frac{2\alpha^{T+1}}{1-\alpha}||c|| \\ &\le 2||c|| \sum_{t=0}^{T}(1-\beta^t) + \frac{2\alpha^{T+1}}{1-\alpha}||c|| \qquad \text{for any } T \ge 1, \end{aligned}$$

which shows that (i) holds.

For (ii), from (2.5) and Lemma 1.2(iii),

$$\lim_{\beta\uparrow 1} |(1-\beta)h_\beta(x,\Phi^{(\infty)}) - \Psi(x,\Phi^{(\infty)},c)| = 0.$$

Also, observing the representation of $h_\beta(x,\Phi^{(\infty)})$ in (2.8), (ii) follows. □

Proof of Theorem 2.1 .

We decompose the occupation measure μ into $\nu_0 \in \mathcal{P}(S)$ and $\Phi \in T(A|S)$ such that (for example,see [1]),

$$\mu(D_1 \times D_2) = \int_{D_1} \Phi(D_2|x)\nu_0(dx) \quad \text{for all } D_1 \in \mathcal{B}_S, D_2 \in \mathcal{B}_A.$$

For this Φ, define $h_\beta(x,\Phi^{(\infty)})$, $h_\beta(x,y)$, $h(x)$ and $h(x,y)$ by (2.8) to (2.11) respectively. Then

$$h_\beta(x,\Phi^{(\infty)}) = \int [c(x,a) + \beta \int h_\beta(y,\Phi^{(\infty)})Q(dy|x,a)]\Phi(da|x) \quad \text{for all } x \in S.$$

So that putting

$$p_\beta(x,a) := c(x,a) + \beta \int h_\beta(y,\Phi^{(\infty)})Q(dy|x,a) - h_\beta(x,\Phi^{(\infty)}),$$

we have

$$\int p_\beta(x,a)\mu(d(x,a)) = 0. \tag{2.12}$$

Also, we have:

$$L^{\pi}_{\nu,\beta}(p_\beta) = L^{\pi}_{\nu,\beta}(c) - L^{\Phi^{(\infty)}}_{\nu,\beta}(c). \tag{2.13}$$

Now, let us prove (2.13). For simplicity, put $h_\beta(x) := h_\beta(x, \Phi^{(\infty)})$. Clearly it holds that, for any given $\pi \in \Pi$,

$$L^{\pi}_{\nu,\beta}(c) = E^{\Phi^{(\infty)}}_{\nu}[h_\beta(X_0)] = \int h_\beta(x)\nu(dx) = E^{\pi}_{\nu}[h_\beta(X_0)]. \tag{2.14}$$

Let

$$W_t := \beta^t c(X_t, \Delta_t) + \beta^{t+1} h_\beta(X_{t+1}) - \beta^t h_\beta(X_t) \quad \text{for each } t \geq 0.$$

Then, since

$$\sum_{t=0}^{T} W_t = \sum_{t=0}^{T} \beta^t c(X_t, \Delta_t) + \beta^{T+1} h_\beta(X_{T+1}) - h_\beta(X_0),$$

we have

$$E^{\pi}_{\nu}[\sum_{t=0}^{T} \beta^t c(X_t, \Delta_t)] - E^{\pi}_{\nu}[h_\beta(X_0)] = E^{\pi}_{\nu}[\sum_{t=0}^{T} W_t] - \beta^{T+1} E^{\pi}_{\nu}[h_\beta(X_{T+1})].$$

As $T \to \infty$ in the above, it follows from (2.14) that

$$L^{\pi}_{\nu,\beta}(c) - L^{\Phi^{(\infty)}}_{\nu,\beta}(c) = E^{\pi}_{\nu}[\sum_{t=0}^{\infty} W_t] = E^{\pi}_{\nu}[\sum_{t=0}^{\infty} E^{\pi}_{\nu}[W_t|\mathcal{B}_t]],$$

where $\mathcal{B}_t = \sigma(X_s, \Delta_s : s \leq t)$. From the definition of W_t, we observe that for each $t \geq 0$,

$$E^{\pi}_{\nu}[W_t|\mathcal{B}_t] = \beta^t p_\beta(X_t, \Delta_t).$$

Therefore, we get (2.13).

Here, we define:

$$p(x,a) := c(x,a) + \int h(y,x)Q(dy|x,a) - \Psi(x, \Phi^{(\infty)}, c). \tag{2.15}$$

Then, the following holds:

Lemma 2.3.

(*i*) $$\lim_{\beta \uparrow 1} ||p_\beta(\cdot,\cdot) - p(\cdot,\cdot)|| = 0.$$

(*ii*) $$\int p(x,a)\mu(d(x,a)) = 0.$$

(*iii*) $$\liminf_{\beta \uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(p_\beta) = \liminf_{\beta \uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(p) \quad \textit{uniformly for } \pi \in \Pi.$$

Proof. For (i),

$$p_\beta(x,a) = c(x,a) + \beta \int h_\beta(y,x)Q(dy|x,a) + (\beta - 1)h_\beta(x, \Phi^{(\infty)}).$$

So:

$$\begin{aligned} |p_\beta(x,a) - p(x,a)| \leq \beta \int |h_\beta(y,x) - h(y,x)|Q(dy|x,a) \\ + (1-\beta)|h(y,x)| + |(1-\beta)h_\beta(x, \Phi^{(\infty)}) - \Psi(x, \Phi^{(\infty)}, c)|. \end{aligned}$$

Thus, from Lemma 2.2, (i) follows. Also, by(2.12) and (i), clearly (ii) holds. Observing (i), for any $\varepsilon > 0$, there exists $\beta_0 < 1$ such that $||p_\beta(\cdot,\cdot) - p(\cdot,\cdot)|| \leq \varepsilon$ for all β with $\beta_0 < \beta < 1$. Therefore, for $\beta(\beta_0 < \beta < 1)$.

$$|L^{\pi}_{\nu,\beta}(p_\beta) - L^{\pi}_{\nu,\beta}(p)| \leq \varepsilon \sum_{t=0}^{\infty} \beta^t = \varepsilon(1-\beta)^{-1},$$

so that

$$|(1-\beta)L^{\pi}_{\nu,\beta}(p_\beta) - (1-\beta)L^{\pi}_{\nu,\beta}(p)| \le \varepsilon,$$

which shows (iii). □

Let us return to the proof of Theorem 2.1. By Lemma 1.2 and (2.4), we get

$$\lim_{\beta\uparrow 1}(1-\beta)L^{\Phi^{(\infty)}}_{\nu,\beta}(c) = \Psi(\nu,\Phi^{(\infty)},c) \quad \text{and} \tag{2.16}$$

$$\liminf_{\beta\uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(c) \le \Psi(\nu,\pi,c). \tag{2.17}$$

It holds from (2.16) that

$$\liminf_{\beta\uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(c) - \Psi(\nu,\Phi^{(\infty)},c) = \liminf_{\beta\uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(p_\beta),$$

which yields by (2.17) and Lemma 2.3(iii) that

$$\Psi(\nu,\pi,c) - \Psi(\nu,\Phi^{(\infty)},c) \ge \liminf_{\beta\uparrow 1}(1-\beta)L^{\pi}_{\nu,\beta}(p_\beta) = L^{\pi}_{\nu}(p). \tag{2.18}$$

By (2.8), (2.10) and Lemma 1.2, it holds

$$\Psi(x,\Phi^{(\infty)},c) = \int h(x)\gamma(dx),$$

so that, observing (2.15), we have

$$p(x,a) = c(x,a) + \int h(y)\overline{Q}(dy|x,a) - h(x).$$

Since $h(x) \le \dfrac{||c||}{1-\alpha}$, $p(x,a) + \dfrac{||c||}{1-\alpha} \ge 0$. Thus, we obtain, from (2.18)

$$\begin{aligned}
&\inf_{\pi\in\Pi'}\Psi(\nu,\pi,c) - \Psi(\nu,\Phi^{(\infty)},c) \\
&\qquad \ge \inf_{\pi\in\Pi'} L^{\pi}_{\nu}(p + \frac{||c||}{1-\alpha}) - \frac{||c||}{1-\alpha} \\
&\qquad \ge K^* \int (p + \frac{||c||}{1-\alpha})\mu(d(x,a)) - \frac{||c||}{1-\alpha} \quad \text{from (2.3)} \\
&\qquad = \frac{(K^*-1)||c||}{1-\alpha} \quad \text{from Lemma 2.3(ii),}
\end{aligned}$$

which completes the proof of Theorem 2.1. □

3. Construction of occupation measures

In this section we construct an occupation measure defined in Section 2 applying the method of obtaining a measure from a pre-measure (see [7] in detail). From this purpose we need a sub-class of policies which is used in the sequel.

For any $(x^o,a^o) \in S \times A$, integer $d \ge 1$ and positive number $\eta > 0$, let

$$\begin{aligned}
\Pi\{x^o,a^o,d,\eta\} := \{\pi \in \Pi \mid P^{\pi}_{x}(X_t = x^o, \Delta_t = a^o \quad &\text{for some } nd \le t < (n+1)d) \ge \eta \\
&\text{for all } n \ge 0 \text{ and } x \in S\}.
\end{aligned}$$

Using a policy π in $\Pi\{x^o,a^o,d,\eta\}$ means the probability that we take action a^o in state x^o during each d-period is no less than η. We define a set function τ on $\mathcal{B}_{S\times A}$ by

$$\tau(D) := \inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} L^{\pi}_{\nu}(D) \quad \text{for all } D \in \mathcal{B}_{S\times A}$$

Then, $\tau(\emptyset) = 0$ and $0 \leq \tau(D) \leq 1$, so τ is a pre-measure. Note that τ satisfies the superadditivity:

$$\tau(\bigcup_{i=1}^{\infty} D_i) \geq \sum_{i=1}^{\infty} \tau(D_i) \tag{3.1}$$

for any sequence $\{D_i\}$ with $D_i \in \mathcal{B}_{S\times A}$ and $D_i \cap D_j = \emptyset (i \neq j)$. Using the pre-measure τ, we define the set function $\tilde{\mu}$ on the collection of all subsets of $S \times A$ by

$$\tilde{\mu}(D) := \sup_{\delta>0} \mu^{\delta}(D), \quad D \subset S \times A, \tag{3.2}$$

where $\mu^{\delta}(D) = \inf_{C_i \in \mathcal{B}_{S\times A}, d(C_i)\leq\delta, D\subset\bigcup C_i} \sum_{i=1}^{\infty} \tau(C_i)$, $d(C) = \sup_{x,y\in C} d(x,y)$, $C \subset S \times A$ and d is a metric on $S \times A$.

Since all Borel sets are $\tilde{\mu}$-measurable, the restriction of $\tilde{\mu}$ to $\mathcal{B}_{S\times A}$ is a measure on $\mathcal{B}_{S\times A}$ (see [7] in detail). We have the following lemma.

Lemma 3.1.

(*i*) $$\tilde{\mu}(D) \leq \tau(D) \quad \text{for all } D \in \mathcal{B}_{S\times A} \quad \text{and}$$

(*ii*) $$\eta/d \leq \tilde{\mu}(S \times A) \leq 1.$$

Proof. Clearly (i) follows from (3.1). For (ii), let $\pi \in \Pi\{x^o, a^o, d, \eta\}$. Then, for any $D \in \mathcal{B}_{S\times A}$ with $(x^o, a^o) \in D$, we have:

$$\begin{aligned}
L_{\nu}^{\pi}(D) &= \liminf_{\beta\uparrow 1}(1-\beta)L_{\nu,\beta}^{\pi}(D) \\
&\geq \liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} P_{\nu}^{\pi}(X_t = x^o, \Delta_t = a^o), \text{ from Lemma 1.2,} \\
&\geq \liminf_{T\to\infty} \frac{1}{T}\sum_{n=0}^{[\frac{T-1}{d}]}\sum_{t=nd}^{(n+1)d-1} P_{\nu}^{\pi}(X_t = x^o, \Delta_t = a^o), \\
&\geq \liminf_{T\to\infty} \frac{1}{T}\sum_{n=0}^{[\frac{T-1}{d}]} P_{\nu}^{\pi}(X_t = x^o, \Delta_t = a^o \text{ for some } nd \leq t < (n+1)d), \\
&\geq \eta/d,
\end{aligned}$$

which implies $\tau(D) \geq \eta/d$ by considering $\pi \in \Pi\{x^o, a^o, d, \eta\}$. Thus, (ii) holds by the definition of $\tilde{\mu}$. □

Now, we can state the main result in this section.

Theorem 3.1. *For $\nu \in P(S)$, there exists an occupation measure associated with $\Pi\{x^o, a^o, d, \eta\}$ and ν.*

Proof. Let $\tilde{\mu}$ be the measure defined by (3.2) with ν. And define $\mu \in P(S \times A)$ by $\mu(\cdot) := \tilde{\mu}(\cdot)/\tilde{\mu}(S \times A)$. From Lemma 3.1(ii), we see μ is well-defined. The proof is completed by showing that μ satisfies the inequality (2.3).

Let g be a non-negative valued simple function defined by

$$g(x) := \sum_{i=1}^{m} \alpha_i I_{D_i}(x),$$

where $D_i \in \mathcal{B}_{S\times A}, D_i \cap D_j = \emptyset (i \neq j), \bigcup_{i=1}^{m} D_i = S \times A$ and $\alpha_i \geq 0 (1 \leq i \leq m)$. Then, we have

$$\inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} L^\pi_\nu(g) = \inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} \liminf_{\beta\uparrow 1} \sum_{i=1}^m \alpha_i(1-\beta)L^\pi_{\nu,\beta}(D_i),$$

from the additivity of $L^\pi_{\nu,\beta}$,

$$\geq \sum_{i=1}^m \alpha_i \tau(D_i) \geq \sum_{i=1}^m \alpha_i \tilde{\mu}(D_i), \quad \text{from Lemma 3.1(i)},$$

$$= K\sum_{i=1}^m \alpha_i \mu(D_i) = K\int g d\mu, \quad \text{where } K = \tilde{\mu}(S\times A).$$

For any $g \in B_+(S\times A)$, let $\{g_n\}$ be a non-decreasing sequence of non-negative real valued simple functions with $\lim_{n\to\infty} g_n = g$. Then by the above result, it holds that

$$K\int g_n d\mu \leq \inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} L^\pi_\nu(g_n) \leq \inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} L^\pi_\nu(g) \quad \text{for all } n \geq 1.$$

As $n \to \infty$ in the above, applying the monotone convergence theorem, we get

$$K\int g d\mu \leq \inf_{\pi\in\Pi\{x^o,a^o,d,\eta\}} L^\pi_\nu(g),$$

which completes the proof. □

References

1. Bertsekas, D. P. and Shreve, S. E., *Stochastic Optimal Control - the Discrete Time Case*, Academic Press, New York, 1978.
2. Blackwell, D. and Ryll-Nardzewski, C., "*Non-existence of every proper conditional distributions*", Ann. Math. Stat., **34**, (1963), pp. 223 - 225.
3. Borkar, V. S., "*A convex analysis approach to Markov decision processes*", Prob. Theory Related Field, **78**, (1988), pp. 583 - 602.
4. Borkar, V. S., *Topics in Controlled Markov Chains*, John Wiley and Sons, Inc, New York, 1991.
5. Hernandez-Lerma, O., *Adaptive Markov Control Processes*, Springer - Verlang, New York, 1989.
6. Kurano, M., "*Markov decision processes with a Borel measurable cost function : the average case*", Math. Oper. Res., **11**, (1986), pp. 309 - 320 .
7. Rogers, C. A. , *Hausdorff Measures*, Cambridge University Press, 1970.
8. Rogers, C. A. , "*Probability measures on compact sets*", Proc. London Math. Soc. (3). **52**, (1986), pp. 328 - 348.
9. Ross, S. M., *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
10. Nummelin, E. , *General Irreducible Markov Chains and Non-negative Operations*, Cambridge univ. Press, 1984.
11. Zygmund, A., *Trigonometric Series*, Cambridge univ. Press, 1968.

Department of Mathematics, Faculty of Science, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba-city, Chiba 263, Japan

E-mail address: msoh@nature.s.chiba-u.ac.jp,mhosaka@nature.s.chiba-u.ac.jp